한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 30, No. 3, P. 51~59

디퓨전 오토인코더의 시선 조작 데이터 증강을 통한 시선 추적

문강륜[1○] 김영한[2] 박용준[3] 김용규[3*]

성균관대학교[1], 가천대학교[2], 비주얼캠프[3]

{kyle[1], rio[2], cody[3], aiden[3]}@visual.camp

Gaze-Manipulated Data Augmentation for Gaze Estimation With Diffusion Autoencoders

Kangryun Moon[1○] Younghan Kim[2] Yongjun Park[3] Yonggyu Kim[3*]

Sungkyunkwan University[1], Gachon University[2], Visualcamp[3]

요 약

시선 벡터 정답값을 갖는 대규모 데이터의 수집은 시선 추적 분야에서 많은 비용을 필요로 한다. 본 논문에서는 원본 사진의 시선을 수정하는 데이터 증강 기법을 사용하여 제한된 개수의 시선 정답값이 주어진 상황에서 시선 추적 모델의 정확도를 향상시키는 방법을 제안한다. 시선 구간 다중 클래스 분류를 보조 작업으로 학습하고, 디퓨전 오토인코더의 잠재 변수를 조정하여 원본 사진의 시선을 편집한 사진을 생성한다. 기존의 얼굴 속성 편집과 달리, 우리는 이진 속성이 아닌 시선 벡터의 피치와 요를 지정한 범주 내로 변경하며, 편집된 사진을 시선 추적 모델의 증강된 학습 데이터로 활용한다. 시선 정답값이 5만 개 이하일 때 준지도 학습에서의 시선 추적 모델의 정확도 향상은 제안한 데이터 증강 기법의 효과를 입증한다.

Abstract

Collecting a dataset with a corresponding labeled gaze vector requires a high cost in the gaze estimation field. In this paper, we suggest a data augmentation of manipulating the gaze of an original image, which improves the accuracy of the gaze estimation model when the number of given gaze labels is restricted. By conducting multi-class gaze bin classification as an auxiliary task and adjusting the latent variable of the diffusion model, the model semantically edits the gaze from the original image. We manipulate a non-binary attribute, pitch and yaw of gaze vector to a desired range and uses the edited image as an augmented train data. The improved gaze accuracy of the gaze estimation network in the semi-supervised learning validates the effectiveness of our data augmentation, especially when the number of gaze labels is 50k or less.

키워드: gaze estimation, data augmentation, diffusion probabilistic model, facial attribute manipulation
Keywords: 시선 추적, 데이터 증강, 디퓨전 확률적 모델, 얼굴 속성 편집

1 서론

시선은 사람의 주의와 의도를 해석하는 데 중요한 단서가 되어 의료 진단[1], 인간과 컴퓨터의 상호작용[2, 3], 증강 현실[4] 등 다양한 영역에서 그 활용 범위를 넓혀가고 있다. 사람의 얼굴이 촬영된 사진으로부터 시선 또는 시선 벡터를 추적하는 방법(appearance-based gaze estimation)은 시선 추적 모델의 학습을 위해 대규모로 수집된 데이터셋을 필요로 하였다[5, 6, 7, 8, 9]. 그러나 시선 벡터가 라벨링된 대규모 데이터셋을 수집하는 것은 많은 비용을 요구하며, 특히 임의의 시선 벡터를 정답값으로 갖는 데이터를 수집할 때에는 제약이 따른다. 데이터 수집 환경에서 실험자마다 임의의 방향을 바라보도록 지시하고, 그때마다 대응되는 정답 시선 벡터를 정확히 계산하기가 어렵기 때문이다. 이러한 제약을 완화하기 위해 원본 얼굴 사진의 시선을 변경하여 사진을 재생성해내는 방법으로 시선 조정(gaze manipulation)이 있다[10]. 눈은 얼굴 사진에서 작은 영역을 차지하고 눈동자

*corresponding author: Yonggyu Kim/Visualcamp(aiden@visual.camp)

Received : 2024.06.15./ Review completed : 1st 2024.07.02. / Accepted : 2024.07.05.
DOI : 10.15701/kcgs.2024.30.3.51
ISSN : 1975-7883(Print)/2383-529X(Online)

의 사소한 이동만으로 시선의 큰 차이를 야기할 수 있기 때문에 원하는 방향을 바라보도록 사진을 조작하는 것은 고난도의 작업이다. 일부 연구에서는 얼굴 사진으로부터 눈 영역만을 분리한 이후 눈 사진의 시선 방향을 조정하려고 시도하였으나[11, 12], 눈 영역을 식별하기 위한 추가적인 전처리 작업이 필요하다는 한계를 가진다.

시선 조정은 얼굴 사진이 가진 고유한 특성 중 하나인 시선을 다룬다는 점에서 얼굴 속성 편집(facial attribute manipulation)이기도 하다. 대부분의 얼굴 속성 편집 연구에서는 적대적 생성 신경망(Generative Adversarial Network)과 같은 생성 모델을 활용하여 웃음의 유무, 머리의 좌우 방향 등 이진 분류가 가능한 얼굴 속성을 편집하였다[13]. 이진 속성 편집은 속성의 발현 유무를 조절할 수 있지만 속성의 발현 정도를 구체적으로 조절할 수 없다는 한계가 있다. 예를 들어 정면을 바라보고 있는 얼굴 사진으로부터 머리가 우측을 향하는 사진으로는 편집할 수 있지만, 명시적으로 머리가 우측 30°만큼 향한 사진으로는 편집할 수 없다. 이에 반해 우리는 특정 시선 방향을 지정하고 원본 사진을 지정한 방향을 바라보는 사진으로 편집하기 위해 시선을 요(yaw)와 피치(pitch) 성분으로 분할하고 각 성분을 범주형 속성으로 간주한다.

본 논문에서는 디퓨전 오토인코더(Diffusion Autoencoder)[14]의 잠재 변수 값을 변경하여 원본 사진으로부터 원하는 방향을 바라보는 사진으로 재생성한다. 그리고 재생성한 사진을 학습 데이터의 증강 데이터로 사용하는 준지도 학습(semi-supervised learning)을 통해 시선 추적 모델의 정확도를 향상시킨다. Figure 1a와 같이 사진의 밝기를 조절하거나 노이즈(noise)를 더하는 일반적인 데이터 증강 기법은 정답값인 시선 벡터를 바꾸지는 않는다. 또한 Figure 1b와 같이 사진 회전 증강으로 시선 벡터를 일부 변형할 수 있지만, 머리의 롤(roll) 방향이 고정된 경우에 대한 시선 벡터의 조정이 어렵다. 본 논문에서는 Figure 1c와 같이 학습 데이터의 정답값인 시선 방향을 다양화하고 정답값의 분포를 고르게 만들어 시선 추적 모델의 정확성과 이상 데이터에 대한 강건성을 향상시킨다.

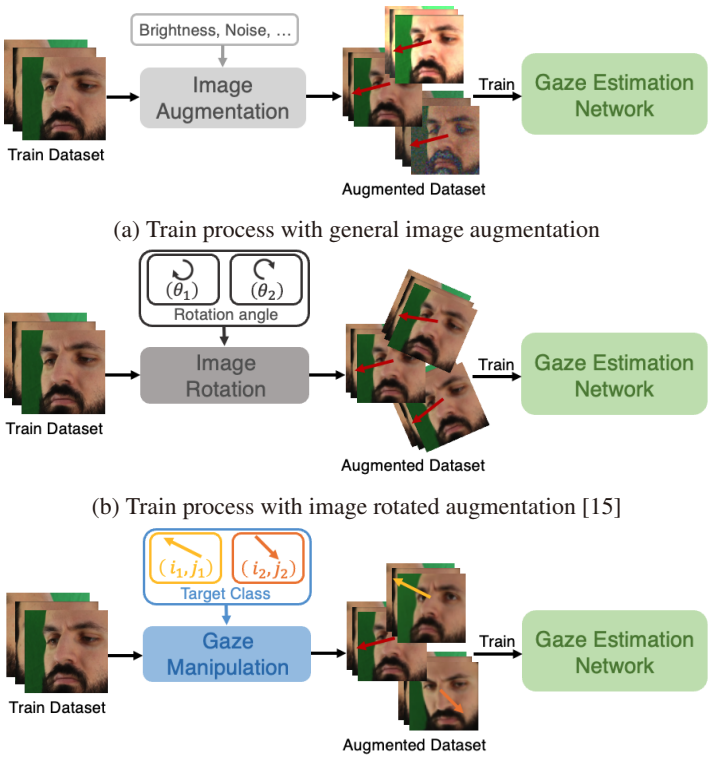


Figure 1: Comparison of general and proposed data augmentation for training a gaze estimation network. Compared to general augmentation, our augmentation revises the gaze label of the original image to a desired class index (i, j), which corresponds to the target angle for pitch and yaw of gaze vector.

2 관련 연구

2.1 시선 추적과 데이터 증강

학습 데이터셋의 시선 정답값과 머리 방향의 넓고 고른 분포는 시선 추적 모델의 정확도와 일반화 능력에 있어서 중요하다[7]. 고르지 못한 시선 정답값의 분포를 갖는 데이터로 학습된 모델은 데이터셋의 범주를 벗어나는 데이터에 대해 내삽 또는 외삽하여 추론해야 하기 때문이다. 일반적으로 학습 데이터셋의 시선 분포를 고르게 하기 위해 사진을 수평으로 뒤집거나 임의의 각도만큼 회전시킨다. RAT[15]는 원본 사진을 회전시키고, 회전 각도만큼 정답 시선 벡터도 변경함으로써 얼굴의 롤 방향 회전에 따른 시선 벡터의 일관성을 학습하고자 하였다. 그러나 시선을 결정하는 중요한 요소인 안구의 움직임을 고려하지는 않았다. 다른 연구자들은 사진을 합성하여 데이터셋의 분포를 다양하게 하고자 하였다. J. Qin과 그의 동료들[16]은 3D 얼굴 복원(3D face reconstruction)을 적용하여 동일한 인물을 다른 시점에서 바라본 사진으로 합성하였다. 그러나 복원 과정에서 얼굴 메시(face mesh)를 사용하기 때문에 사진에서 얼굴을 벗어나는 영역은 현실 세계의 데이터와 이질적인 부분이 존재한다.

2.2 디퓨전 모델

디퓨전 확률적 모델(Diffusion Probabilistic Model)은 사진 생성[17, 18], 사진 인페인팅(image inpainting)[19], 사진 편집[20] 등 다양한 생성 작업 분야에서 뛰어난 결과를 보여주면서 사진 합성 분야의 새로운 접근법으로 부상하고 있다. 디퓨전 확률 모델은 사진에 더해지는 노이즈(noise)를 예측하여 궁극적으로 특정 분포에서 추출(sample)된 임의의 노이즈로부터 본래의 사진을 재생성한다. DDPM[21]은 마르코프 체인(Markov chain)을 적용하며 각 타임스탬프(timestamp)마다 가우시안 분포(Gaussian distribution)에서 추출한 노이즈를 원본 사진에 점진적으로 더하고, 해당 단계에서 더해진 노이즈를 학습함으로써 고품질의 사진을 생성하였다. 그러나 노이즈 제거를 위해 타임스탬프의 수만큼 반복 연산을 수행하기 때문에 사진 재생성에 상대적으로 긴 시간이 소

요된다는 한계가 있다. 이 문제를 완화하기 위해 DDIM[22]은 각 노이즈 제거 단계를 결정론적으로 강제함으로써 사진 재생성에 필요한 단계 수를 줄였다. 확률적인 요소의 개입을 배제함으로써 사진 재생성의 소요 시간이 단축되었을 뿐만 아니라 한 입력값에 대해서 일관되게 재현되는 사진을 얻을 수 있었다. Diff-AE[14]에서는 디퓨전 모델의 잠재 변수를 변화시켜 사진을 편집하여 디퓨전 모델에서 잠재 공간의 해석이 가능함을 보였다. Diff-Video-AE[23]는 사전 학습된 인코더(encoder)를 사용하여 잠재 변수를 추출하고 사진의 얼굴 속성을 편집하였다.

2.3 얼굴 속성 편집

얼굴 속성 편집은 변경하고자 하는 얼굴 속성 외의 다른 속성들은 유지하면서 원본 사진을 편집하는 작업이다. 얼굴 속성은 그 유형에 따라 이진형, 범주형, 또는 연속형으로 나누어지며 편집 과정에서 하나 또는 여러 개의 속성을 조절할 수 있다. 구체적인 구현 방법으로는 사진 변환(image translation), 사진 세그멘테이션(image segmentation), 또는 생성형 모델의 잠재 공간(latent space) 해석이 있다[24, 25]. 최근에는 생성형 모델의 잠재 공간 상에 존재하는 잠재 벡터(latent vector)를 수정하여 사진을 편집하려는 시도들이 연구되고 있다[26].

StyleGAN[27]은 각 사진에 대응되는 잠재 벡터를 보간하여 서로 다른 두 스타일의 사진 간의 합성 가능성을 확인하였다. InterFaceGAN[13]은 이진 속성을 구별할 수 있는 경계(boundary)가 존재하는 경우 해당 속성의 발현 유무를 조절할 수 있음을 이진형 얼굴 속성 편집을 통해 검증하였다. Diff-AE[14]는 디퓨전 모델의 잠재 변수를 추가로 정의하여 적대적 생성형 모델과 마찬가지로 잠재 변수의 보간을 통해 얼굴 속성 편집이 가능함을 보였다. Diff-Video-AE[23]는 사전 학습된 인코더를 사용해 연속된 비디오 프레임에서 얼굴 속성을 조작하였다.

3 사전 지식

3.1 생성형 모델의 잠재 공간 해석과 사진 편집

생성형 모델의 잠재 공간에 존재하는 벡터를 사전 학습된 선형 분류기의 가중치 기울기(gradient) 방향으로 이동시키면 원본 사진 속 이진 속성의 발현 정도를 조절할 수 있다[28, 13, 14]. 잠재 벡터 $\mathbf{z}$와 $\mathbf{z}$를 입력으로 받아 특정 이진 속성의 존재 유무를 학습한 선형 분류기(linear classifier) $\mathcal{C}$로부터 식 1과 같이 벡터 $\mathbf{z}$를 $\mathbf{z}'$으로 변형할 수 있다.

$$\mathbf{z}' = \mathrm{L2Norm}(\mathbf{z} + s\mathbf{w})\,. \tag{1}$$

이때 $\mathbf{w}$는 선형 분류기의 가중치 벡터, s는 발현 정도를 조절하는 상수이다. 예를 들어 선형 분류 계층이 잠재 벡터로부터 입력 사진의 사람의 머리가 오른쪽을 향하는지의 유무를 판단하도록 학습했다면, s의 값이 증가함에 따라 $\mathbf{z}'$으로부터 재생성된 사진은 다른 얼굴 속성은 유지한 상태에서 점차 오른쪽을 바라보도록 편집된다. 본 논문에서는 해당 개념을 범주형 속성으로 확장한다.

4 제안하는 방법

4.1 개요

제안하는 프레임워크의 목표는 원본 사진으로부터 사람의 시선 방향을 조정한 사진을 생성해내는 것이며 이를 위해 학습 단계와 조작 단계로 구분한다. 학습 단계에서는 사진을 재현하고, 시선과 관련된 특징과 사진 편집에 용이한 특징을 담은 잠재 변수 $\mathbf{z}_{\mathrm{face}}$를 추출한다(Figure 2a-(1)). 사전 학습된 시선 인코더(gaze encoder) E_{gaze}와 신원 인코더(identity encoder) E_{id}는 입력 사진 $\mathbf{x}_0$를 $E_{gaze}(\mathbf{x}_0) = \mathbf{z}_{\mathrm{gaze}}$와 $E_{id}(\mathbf{x}_0) = \mathbf{z}_{\mathrm{id}}$의 잠재 벡터로 각각 임베딩한다. 두 벡터는 식 2와 같이 직렬로 연결된 뒤 선형 계층을 통과하여 잠재 변수 $\mathbf{z}_{\mathrm{face}}$를 얻는다. 두 선형 분류기 $\mathcal{C}_{pitch}$와 $\mathcal{C}_{yaw}$는 식 $\mathcal{C}_{pitch}(\mathbf{z}_{\mathrm{face}}) = \hat{y}_{pitch}$와 식 $\mathcal{C}_{yaw}(\mathbf{z}_{\mathrm{face}}) = \hat{y}_{yaw}$을 따라 잠재 변수 $\mathbf{z}_{\mathrm{face}}$로부터 입력 사진의 시선이 속할 구간을 예측한다(Figure 2a-(2)). 조건부 DDIM 디코더는 $\mathbf{z}_{\mathrm{face}}$를 조건부 입력으로 받아 잠재 변수 $\mathbf{x}_{\mathrm{T}}$에서 노이즈를 제거한 $\hat{\mathbf{x}}_0$를 생성한다(Figure 2a-(3)). 조작 단계에서 조건부 DDIM 디코더는 $\mathbf{z}_{\mathrm{face}}$ 대신 의미론적 시선 조작 모듈(Semantic Gaze Manipulation Module, SGMM)을 거쳐 편집된 $\mathbf{z}'_{\mathrm{face}}$를 조건부 입력값으로 받아 원본 사진으로부터 시선이 조정된 $\hat{\mathbf{x}}'_0$를 생성한다. Figure 2a는 학습 단계에서 모델의 구조를, Figure 2b는 조작 단계에서의 의미론적 시선 조작 모듈을 나타낸 것이다.

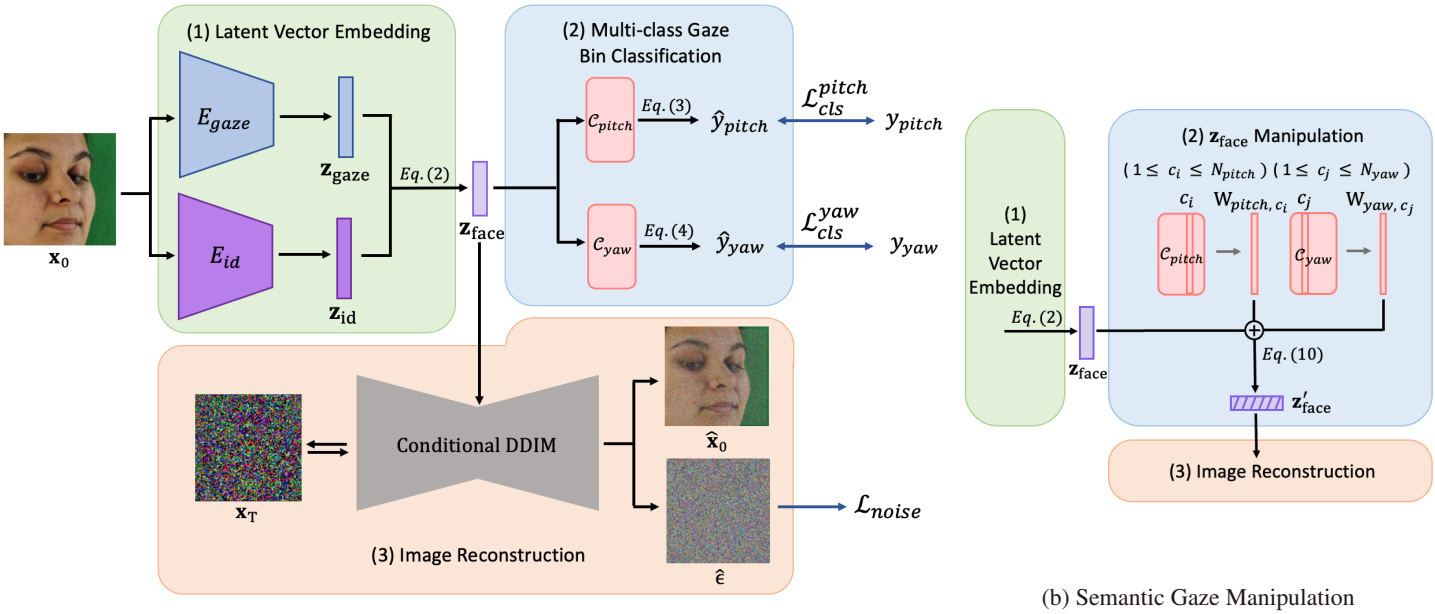
4.2 사전 학습된 인코더와 선형 분류기

인코더의 목표는 원본 사진과 동일한 사진을 재현하면서도 조작 단계에서 시선을 조절할 때 필요한 잠재 변수 $\mathbf{z}_{\mathrm{face}}$를 추출하는 것이다. 사전 학습된 시선 추적 모델[7]은 시선과 관련 특징을 제공할 $\mathbf{z}_{\mathrm{gaze}}$로 인코딩한다. ArcFace[29]는 사람의 얼굴을 분류하도록 학습되었기 때문에 다른 특징 추출기(feature extractor)에 비해 시선 각도가 변화하여도 동일한 사람에 대해서는 일관된 인코딩 벡터 $\mathbf{z}_{\mathrm{id}}$를 제공한다. 인코딩된 두 벡터는 식 2에 따라 직렬로 연결된 뒤 선형 계층을 통과하여 최종적으로 $\mathbf{z}_{\mathrm{face}}$가 된다. $\mathbf{z}_{\mathrm{face}}$는 두 개의 선형 분류기와 조건부 DDIM에 모두 전달된다.

선형 분류기의 목표는 범주형 시선 분류 작업을 학습하여 조작 단계에서 식 1과 같이 각 시선 클래스에 대응되는 선형 분류기의 가중치를 얻는 것이다. 우리는 시선의 피치와 요를 범주형 속성으로 간주하여 원하는 방향을 바라보는 사진을 생성한다. 두 개의 선형 분류기는 다음과 같이 정의한다:

$$\mathbf{z}_{\mathrm{face}} = \mathrm{MLP}([\mathbf{z}_{\mathrm{gaze}}; \mathbf{z}_{\mathrm{id}}])\,, \tag{2}$$

$$\mathcal{C}_{pitch}(\mathbf{z}_{\mathrm{face}}) = \mathrm{Softmax}(\mathrm{W}_{\mathrm{pitch}}{}^{\intercal}\mathbf{z}_{\mathrm{face}} + b_1)\,, \tag{3}$$



(a) Overview of architecture at train phase

(b) Semantic Gaze Manipulation Module(SGMM) for image manipulation phase

Figure 2: Model overview. At the train phase, the model learns to reconstruct images from the original image $\mathbf{x}_0$ and latent variable $\mathbf{z}_{\text{face}}$. Meanwhile, the predicted probability for each class $\hat{y}_{pitch}$ and $\hat{y}_{yaw}$ from the linear classifier is compared with the ground truth label y_{pitch} and y_{yaw}. At the manipulation phase, our Semantic Gaze Manipulating Module(SGMM) aims to generate a revised image of a person looking in the direction of a target pitch and yaw class, c_i and c_j. The weight vector W_{pitch,c_i} and W_{yaw,c_j} are combined with $\mathbf{z}_{\text{face}}$ to generate $\mathbf{z}'_{\text{face}}$. According to the phase, conditional input either $\mathbf{z}_{\text{face}}$ or $\mathbf{z}'_{\text{face}}$ is passed to the conditional DDIM with timestamp t and image $\mathbf{x}_t$, which predicts added noise $\hat{\epsilon}$.

$$\mathcal{C}_{yaw}(\mathbf{z}_{\text{face}}) = \text{Softmax}(\mathrm{W}_{\text{yaw}}{}^{\intercal}\mathbf{z}_{\text{face}} + b_2)\ , \tag{4}$$

W_{pitch}와 W_{yaw}는 선형 분류기의 가중치 행렬을, b_1, b_2는 편향(bias)을 나타낸다. $\mathcal{C}_{yaw}$가 $[-\theta, \theta]$의 범주에 대해 잠재 변수 $\mathbf{z}_{\text{face}}$가 어느 구간에 속하는지 N_{yaw}개의 클래스로 분류하는 경우를 가정해보자. 이때 $\mathcal{C}_{yaw}$는 입력값 $\mathbf{z}_{\text{face}}$로부터 입력 사진에 대한 정답 시선 벡터의 요가 각 시선 구간 클래스에 대응될 확률을 예측한다. 생성형 모델의 잠재 공간을 이용하는 일반적인 얼굴 속성 편집 모델과 달리 선형 분류기를 훈련 과정에 포함시킴으로써 시선 정보와 더 밀접한 관계를 갖는 잠재 변수 $\mathbf{z}_{\text{face}}$를 추출한다.

4.3 학습 손실 함수

4.3.1 노이즈 예측 손실 함수

조건부 DDIM[22] ϵ_θ는 타임스템프 $t(0 < t \leqslant T)$에서의 사진 $\mathbf{x}_t$, 잠재 변수 $\mathbf{z}_{\text{face}}$와 타임스템프 t로부터 원본 사진 $\mathbf{x}_0$에 더해진 노이즈 ϵ_t를 학습한다. $\mathbf{z}_{\text{face}}$는 조건부 DDIM이 시선과 신원의 특징을 고려하여 노이즈를 예측하도록 돕는다. 노이즈 예측에 대한 손실 함수 $\mathcal{L}_{noise}$는 L2 손실 함수이며 다음과 같이 정의한다:

$$\mathcal{L}_{noise} = \mathbb{E}_{\mathbf{x}_0,\epsilon_t}||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{z}_{\text{face}}) - \epsilon_t||_2^2\ , \epsilon_t \sim \mathcal{N}(0,\ I)\ . \tag{5}$$

4.3.2 분류 손실 함수

잠재 변수 $\mathbf{z}_{\text{face}}$가 시선 관련 특징을 추출하고 사진 편집 단계에서 사용될 가중치 벡터를 얻기 위해 다중 클래스 분류를 수행한다. 각 선형 분류기는 잠재 변수 $\mathbf{z}_{\text{face}}$를 입력받아 잠재 벡터로부터 입력 사진 시선 벡터의 요와 피치를 구간 단위로 분류하도록 (multi-class gaze bin classification) 학습한다[30]. 이를 통해 조작 단계에서 입력 사진의 시선 벡터가 달라지는 경우 $\mathbf{z}_{\text{face}}$가 알맞게 변화하여 조건부 DDIM이 목표하는 시선 방향을 가진 사진을 생성하도록 보조한다. 시선의 요와 피치의 구간 다중 클래스 분류 손실 함수는 다중 클래스 교차 엔트로피 손실 함수(multi-class cross-entropy loss)이며 다음과 같이 정의한다:

$$\mathcal{L}_{cls}^{pitch} = -\sum_{c=1}^{N_{pitch}} w_c^{pitch} \cdot y_c^{pitch} \cdot \log(\hat{y}_c^{pitch})\ , \tag{6}$$

$$\mathcal{L}_{cls}^{yaw} = -\sum_{c=1}^{N_{yaw}} w_c^{yaw} \cdot y_c^{yaw} \cdot \log(\hat{y}_c^{yaw})\ , \tag{7}$$

$$\mathcal{L}_{cls} = \mathcal{L}_{cls}^{pitch} + \mathcal{L}_{cls}^{yaw}\ , \tag{8}$$

w_c^{pitch}, w_c^{yaw}는 각각 피치와 요의 목표 클래스 c에 대한 가중치, N_{pitch}, N_{yaw}는 클래스 수, y_c^{pitch}, y_c^{yaw}는 클래스 정답값, $\hat{y}_c^{pitch}$, $\hat{y}_c^{yaw}$는 각 클래스에 속할 확률 예측값이다. 클래스별 가중치는 데이터셋을 구성하는 클래스 간 사진 수의 불균형을 고려

한다. 전체 손실 함수 $\mathcal{L}_{total}$ 은 다음과 같다:

$$\mathcal{L}_{total} = \mathcal{L}_{noise} + \gamma \mathcal{L}_{cls} \ , \tag{9}$$

γ는 사진 재현 작업과 분류 작업간의 균형을 맞추기 위한 계수이다.

4.4 사진 조작

사진 조작 단계에서는 원본 사진으로부터 목표하는 시선의 피치와 요를 바라보는 사진으로 편집한다. 이를 위해서는 식 1과 같이 목표 피치 또는 요 구간에 대응되는 가중치 벡터가 필요하다. 피치 선형 분류기와 요 선형 분류기의 가중치 행렬 $\mathrm{W}_{pitch}, \mathrm{W}_{yaw}$는 각각 $(d,\, N_{pitch}), (d,\, N_{yaw})$의 차원을 가지며, d는 잠재 변수 $\mathbf{z}_{\mathrm{face}}$의 차원이다. Figure 2b와 같이 목표하는 피치와 요에 대응되는 클래스 번호가 c_i, $c_j (1 \leqslant c_i \leqslant N_{pitch},\, 1 \leqslant c_j \leqslant N_{yaw})$일 때, 가중치 벡터는 각각 W_{pitch}의 c_i번째 열벡터와 W_{yaw}의 c_j번째 열벡터이다. 잠재 변수 $\mathbf{z}_{\mathrm{face}}$는 식 10과 같이 선형 분류기 가중치의 기울기 방향으로 이동한 $\mathbf{z}'_{\mathrm{face}}$로 변형된다. 이때 특정 속성의 변형이 과도하게 강조되지 않도록 L2 정규화(L2 normalization)를 적용한다.

$$\mathbf{z}'_{\mathrm{face}} = \mathrm{L2Norm}(\mathbf{z}_{\mathrm{face}} + s(\mathrm{W}_{pitch,c_i} + \mathrm{W}_{yaw,c_j})) \, . \tag{10}$$

s는 조작의 정도를 조절하는 상수이며, s가 클수록 조작의 정도가 증가한다. Figure 3은 편집된 사진들의 예시이다.

5 실험

5.1 데이터셋 및 구현 방법

시선 조작 증강 데이터셋을 활용한 모델의 정확도 향상을 평가하기 위해 ETH-XGaze 데이터셋[7]을 사용한다. ETH-XGaze 데이터셋은 110명의 실험자로부터 수집한 110만 장의 공개 데이터셋이며, 훈련 데이터셋 80명의 인원 중 5명을 분할하여 검증 데이터셋으로 사용하였다. 시선 벡터가 라벨링된 학습 데이터셋의 수에 따른 시선 추적 모델의 정확도 향상 정도를 비교하기 위해 각 피실험자에 대해 동일한 개수의 사진을 무작위로 추출하여 3k, 5k, 10k, 20k, 50k개의 사진과 시선 정답값으로 구성된 5가지의 하위 데이터셋을 선별하였다.

ETH-XGaze 데이터셋 시선 정답값이 가지는 피치의 범위는 $[-70°, 70°]$이고 요의 범위는 $[-120°, 120°]$이다. 피치와 요의 구간 분류를 위한 클래스 개수 N_{pitch}와 N_{yaw}는 각각 14와 24이며, 그 결과 하나의 클래스는 10°의 구간에 속하는 각도들을 포함한다. 잠재 벡터 $\mathbf{z}_{\mathrm{gaze}}, \mathbf{z}_{\mathrm{id}}$와 잠재 변수 $\mathbf{z}_{\mathrm{face}}$의 차원 d는 모두 512이며, 계수 γ는 0.001로 설정하였다. 각 사진의 크기는 128x128로 조정하였고 정규화 과정[31]을 따른다. 사진 조작을 위한 스케일 계수 s는 0.3이다. 모델은 NVIDIA H100 GPU를 사용해 1,500만 번의 반복(iteration)으로 학습하였다.

Table 1: Evaluation of semi-supervised learning with respect to number of sub-dataset. Each value denotes the angular error between the predicted and ground truth gaze vector in degrees.

Method	3k	5k	10k	20k	50k
Baseline	10.69	8.89	7.42	6.47	5.60
RAT [15]	9.47	8.25	6.83	6.34	5.89
Ours	**9.40**	**8.20**	**6.68**	**6.09**	**5.48**

5.2 시선 조작 증강 기법의 평가 결과

시선 정답값을 수정하는 증강 기법의 효과를 검증하기 위해 다른 개수의 데이터로 구성된 5개의 하위 데이터셋에서 준지도 학습을 수행하였다. 시선 인코더 E_{gaze}는 시선 레이블이 있는 하위 데이터셋으로만 사전 학습하였다. 학습 단계에서 조건부 DDIM은 ETH-XGaze 훈련 데이터셋의 사진을 시선 정답값 없이 모두 사용한다. 반면 시선 구간 분류에 대한 학습은 시선 정답값이 있는 하위 데이터셋에 대해서만 이루어진다. 따라서 노이즈 예측 손실은 매번 조건부 DDIM에 전파되지만, 분류 손실은 하위 데이터셋의 경우에만 선형 분류기에 전파된다. 조작 단계에서는 하위 데이터셋과 편집된 사진의 개수가 같도록 생성하였다. 편집된 사진에 대한 수도 정답값(pseudo label)은 하위 데이터셋으로 학습한 시선 추적 모델로부터 계산되었다. 최종적으로는 하위 데이터셋과 증강된 편집 데이터셋을 모두 사용하여 준지도 학습하였다. 공정한 비교를 위해 특징 추출 모델(backbone)은 ResNet-50[32]으로 통일하였다. 좌우 반전된 이미지 증강을 적용하고 하위 데이터셋만으로 학습한 경우를 베이스라인(baseline)으로 선정하고, 다른 증강 기법을 적용한 경우의 시선 추적 모델과의 정확도를 비교하였다. 모델의 정확도는 각도 오차 θ_{error}로 평가하며 예측 시선 단위벡터 $\hat{\mathbf{g}}$ 와 정답 시선 단위벡터 $\mathbf{g}$의 각도 오차는 식 11과 같다.

$$\theta_{error} = \arccos(\hat{\mathbf{g}} \cdot \mathbf{g}) \, . \tag{11}$$

우리는 회전된 사진에 따라 시선 벡터를 조절하는 RAT[15]를 비교군으로 한다. 사진은 $[-30°, 30°]$ 내의 임의의 각도로 회전시켰으며, 회전 각도 θ에 대해 정답값 시선 벡터 $\mathbf{g}$는 $\mathbf{g}' = \mathrm{R}\mathbf{g}^{\intercal}$로 변경하였다. 이때 R는 θ에서 파생된 회전 행렬이다. 공정한 비교를 위해 동일한 수의 데이터를 증강하였다. Table 1에서와 같이 우리의 방법은 5가지의 경우에 대해 베이스라인과 비교군보다 높은 정확성을 보였으며, 베이스라인보다는 평균적으로 7.56%의 정확도가 향상되었다. 특히 하위 데이터셋의 크기가 5만 개일 때 회전 증강의 경우 성능이 저하되는 반면 우리의 방법은 성능이 향상되었다. 사진 회전 증강은 머리의 롤 방향 회전에 따른 시선의 일관성은 학습할 수 있지만 안구의 움직임에 따른 시선은 학습하기 어렵기 때문이다. 따라서 선별된 데이터의 수가 증가함에 따라 증강의 효과는 감소하고, 변형한 시선 벡터의 부정확함으로 인해 베이스라인보다 모델의 성능 개선이 저하될 수 있다.

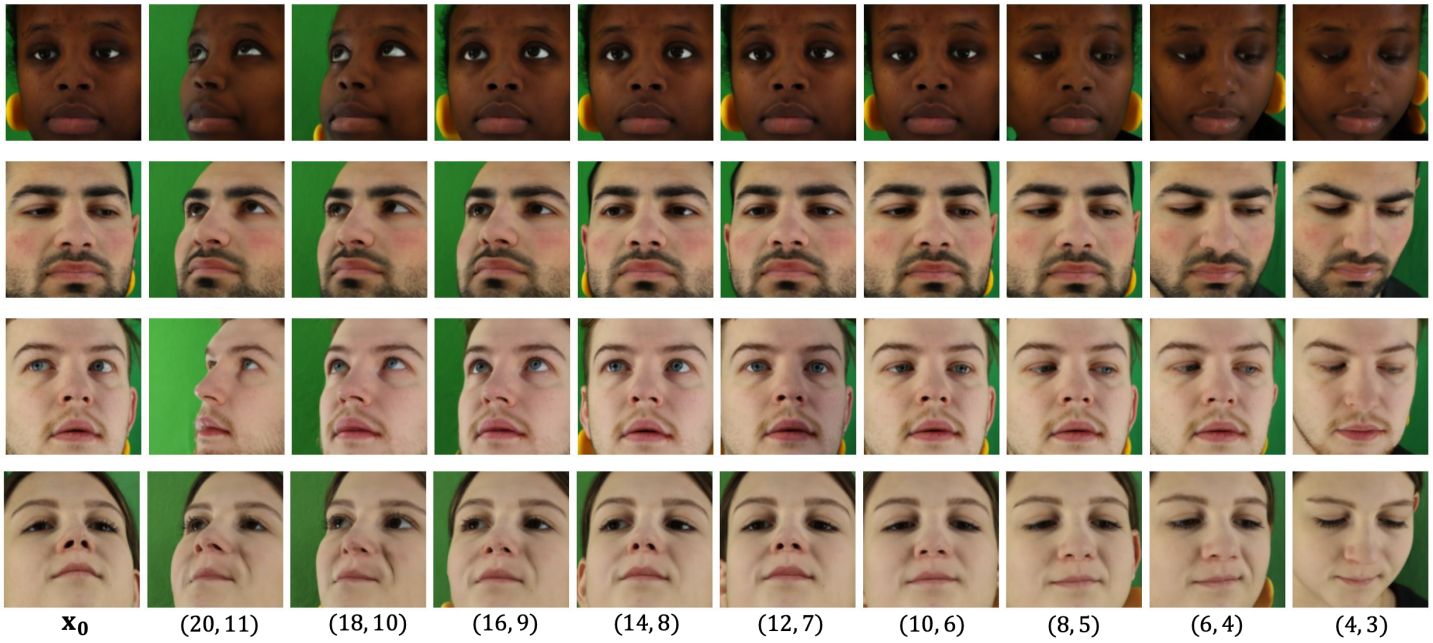


Figure 3: Visualization of gaze-manipulated augmented images. The first and second integers in each parentheses refer to the target class index for the yaw and pitch of a gaze vector. The gaze vector of the revised image moves leftwards as the yaw index decreases and downwards as the pitch index decreases, in the perspective of the subject.

Table 2: Ablation study on auxiliary task, multi-class gaze bin classification. Each value denotes the accuracy of multi-class gaze bin classification.

Ablation on ↓	Details		Accuracy (%)	
	joint.	multi.	pitch	yaw
Binary	✓	✗	33.05	12.94
Categorical	✗	✓	34.92	15.48
Ours	✓	✓	**35.27**	**16.16**

5.3 비교 실험

본 절에서는 시선의 요와 피치의 범주형 속성 해석과 선형 분류기의 공동 학습(joint training)이 시선 구간 분류 보조 학습에 미치는 영향을 비교 실험을 통해 분석한다. 이를 통해 잠재 변수 $\mathbf{z}_{\text{face}}$에 시선과 관련 속성이 추출된 정도를 간접적으로 확인한다. k 번째 데이터에 대한 시선 클래스의 정답값과 예측 확률값을 y_k와 $\hat{y}_k$, 검증 데이터의 수를 n이라할 때, 보조 학습의 정확도(Accuracy)는 식 12와 같다.

$$\text{Accuracy} = \frac{\sum_{k=1}^{n} \mathbf{1}(y_k = \arg\max(\hat{y}_k))}{n}. \tag{12}$$

Table 2는 비교 실험의 결과이다. 첫 번째 행(Binary)은 선형 분류기가 학습 단계에 포함(joint.)되지만 범주형 속성이 아닌 이진 속성으로 분류한 경우를, 두 번째 행(Categorical)은 시선의 피치와 요를 다중 클래스로 분류(multi.)하지만 시선 범주 분류 학습을 확산 모델의 학습 이후에 별도로 진행한 경우를 의미한다. 잠재 변수 $\mathbf{z}_{\text{face}}$가 시선 정보와 무관하게 균등 분포(uniform distribution)를 따른다면 피치의 경우 약 7%, 요의 경우 약 4%의 정확도를 가져야 한다. 그러나 세 가지의 경우에서 모두 균등 분포일 때보다 높은 정확도를 보였다는 점에서 잠재 변수가 시선과 관련된 특징을 보유하고 있음을 알 수 있다. 시선을 이진 속성으로 분류한 경우 보조 작업에서 가장 낮은 성능을 보였는데, 이는 개선된 표현 학습(representation learning)을 위해서는 시선의 범주를 세분화해야 한다는 것을 의미한다. 다만 데이터셋에서 요의 정답값 범주가 피치 정답값 범주보다 넓기 때문에 세 가지 경우에서 모두 요의 예측 정확도가 피치의 예측 정확도보다 낮은 경향을 보인다. 제안한 방법처럼 공동 훈련과 시선을 범주형 속성으로 학습한 경우에 시선을 분류할 때 가장 높은 정확도를 보였다.

Figure 4: Visualization of original images and manipulated images from unseen datasets[6, 9] and limitations. 'Orig.' denotes input image $\mathbf{x_0}$ and 'Manip.' denotes gaze-manipulated image.

6 결론 및 향후 연구

본 논문에서는 시선 추적을 위해 원본 사진의 시선을 조작하는 새로운 증강 기법을 제시했다. 시선 벡터가 라벨링된 5만 개 이내의 하위 데이터셋과 라벨링되지 않은 데이터셋으로 사진 상의 얼굴이 원하는 방향을 바라보도록 편집할 수 있었으며, 다른 데이터 증강 기법에 비해 더 큰 폭의 정확도 향상을 야기했다. 또한

디퓨전 모델을 사용하여 범주형 속성을 편집한 방법은 기존 이진형 속성을 중점적으로 다룬 얼굴 편집 작업에 다양하게 활용될 수 있을 것으로 기대된다.

본 논문에서 제안한 방법은 편집하고자 하는 사진이 학습 도메인을 벗어나 있거나 목표하는 시선이 원본 사진의 시선과 큰 폭으로 벗어난 경우 Figure 4의 첫 번째와 두 번째 예시와 같이 시선 편집에 어려움을 겪을 수 있다. 또한 Figure 4의 세 번째 예시와 같이 시선 편집 과정에서 인물의 정체성이 동일하게 유지되지 않는 경우도 존재한다. 디퓨전 모델의 잠재 공간 해석이 학습 도메인을 벗어난 모든 사진에 일관되게 적용되기에는 어려운 부분이 존재하기 때문이다. 이는 디퓨전 모델의 잠재 공간에 대한 추가적인 연구를 통해 사진 편집의 일반화 능력을 보완할 수 있을 것이다.

감사의 글

이 논문은 2024년도 비주얼캠프의 지원을 받아 수행된 연구임

References

[1] S. De Silva, S. Dayarathna, G. Ariyarathne, D. Meedeniya, S. Jayarathna, and A. M. Michalek, "Computational decision support system for adhd identification," *International Journal of Automation and Computing*, vol. 18, no. 2, pp. 233–255, 2021.

[2] J. H. Goldberg, M. J. Stimson, M. Lewenstein, N. Scott, and A. M. Wichansky, "Eye tracking in web search tasks: design implications," in *Proceedings of the 2002 symposium on Eye tracking research & applications*, 2002, pp. 51–58.

[3] S. Hong, Y. Kim, and T. Park, ""blinks in the dark": Blink estimation with domain adversarial training (beat) network," *IEEE Transactions on Consumer Electronics*, 2023.

[4] J.-Y. Lee, H.-M. Park, S.-H. Lee, T.-E. Kim, and J.-S. Choi, "Design and implementation of an augmented reality system using gaze interaction," in *2011 International Conference on Information Science and Applications*. IEEE, 2011, pp. 1–8.

[5] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE and the Computer Vision Foundation, 2016, pp. 2176–2184.

[6] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Mpiigaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 162–175, 2017.

[7] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, "Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 365–381.

[8] S. H. Choi, D. Son, Y. Ha, Y. Kim, S. Hong, and T. Park, "Looking to personalize gaze estimation using transformers," *Journal of Computing Science and Engineering*, vol. 17, no. 2, pp. 41–50, 2023.

[9] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras," in *Proceedings of the symposium on eye tracking research and applications*, 2014, pp. 255–258.

[10] Y. Ganin, D. Kononenko, D. Sungatullina, and V. Lempitsky, "Deepwarp: Photorealistic image resynthesis for gaze manipulation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 311–326.

[11] Y. Yu, G. Liu, and J.-M. Odobez, "Improving few-shot user-specific gaze adaptation via gaze redirection synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 937–11 946.

[12] K. Wang, R. Zhao, and Q. Ji, "A hierarchical generative model for eye image synthesis and eye gaze estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 440–448.

[13] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9243–9252.

[14] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, "Diffusion autoencoders: Toward a meaningful and decodable representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 619–10 629.

[15] Y. Bao, Y. Liu, H. Wang, and F. Lu, "Generalizing gaze estimation with rotation consistency," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE and the Computer Vision Foundation, 2022, pp. 4207–4216.

[16] J. Qin, T. Shimoyama, and Y. Sugano, "Learning-by-novel-view-synthesis for full-face appearance-based 3d gaze estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4981–4991.

[17] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[19] G. Batzolis, J. Stanczuk, C.-B. Schönlieb, and C. Etmann, "Conditional image generation with score-based diffusion models," *arXiv preprint arXiv:2111.13606*, 2021.

[20] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "Ilvr: Conditioning method for denoising diffusion probabilistic models," 2021.

[21] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[22] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[23] G. Kim, H. Shim, H. Kim, Y. Choi, J. Kim, and E. Yang, "Diffusion video autoencoders: Toward temporally consistent face video editing via disentangled video encoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6091–6100.

[24] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[25] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 1857–1865.

[26] A. Nickabadi, M. S. Fard, N. M. Farid, and N. Mohammadbagheri, "A comprehensive survey on semantic facial attribute editing using generative adversarial networks," *arXiv preprint arXiv:2205.10587*, 2022.

[27] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[28] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang, "Gan inversion: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 3, pp. 3121–3138, 2022.

[29] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE and the Computer Vision Foundation, 2019, pp. 4690–4699.

[30] A. A. Abdelrahman, T. Hempel, A. Khalifa, A. Al-Hamadi, and L. Dinges, "L2cs-net : Fine-grained gaze estimation in unconstrained environments," in *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*. IEEE, 2023, pp. 98–102.

[31] X. Zhang, Y. Sugano, and A. Bulling, "Revisiting data normalization for appearance-based gaze estimation," in *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, 2018, pp. 1–9.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

〈저자소개〉

문강륜
- 성균관대학교 재학 2019년 ~ 현재
- 2022년 9월 ~ 현재 비주얼캠프 연구팀 소속
- 관심 분야 : 시선 예측, 이미지 생성
- https://orcid.org/0009-0005-2557-470X

김영한
- 2023년 1월 ~ 현재 비주얼캠프 연구팀 소속
- 관심분야 : 시선 예측, 인간-물체 상호작용 인식
- https://orcid.org/0009-0004-5145-2100

박용준
- 2024년 1월 ~ 현재 비주얼캠프 연구팀 소속
- 관심분야 : 시선 예측, 객체 탐지, 자기 지도 학습
- https://orcid.org/0000-0002-2261-0956

김용규
- 2019년 3월 한국기술교육대학교 컴퓨터공학과 졸업(석사)
- 2021년 8월 ~ 현재 비주얼캠프 연구팀
- 관심분야 : 컴퓨터비전, 시선 예측, 3차원 형상 복원
- https://orcid.org/0000-0001-7038-8715